Primarily for non-mathematicians.

SOME COMPUTER USES IN STATISTICS AND GENETICS *

BU-151-M S. R. Searle November, 1962

ABSTRACT

Several uses of a computer for work in statistics and genetics are described by means of examples related to agricultural research. They consist of problems in both data processing and simulation techniques.

*Biometrics Unit, Plant Breeding Department, Cornell University

SOME COMPUTER USES IN STATISTICS AND GENETICS *

BU-151-M S. R. Searle November, 1962

INTRODUCTION

High-speed computers are helpful in all fields of work not only because they speed up already-existing operations but also because they make feasible undertakings that are otherwise excessively arduous. This is as true in statistics and genetics as it is in other fields of endeavor. Just as the businessman by using a computer can have his accounts summarized in numerous ways in less time than was previously required for a single summary, so the statistician can now undertake the calculations for many experiments and surveys in the time once needed for just one or two. This paper outlines some examples of the statistician's work which illustrate the advantages of computer speed both in accomplishing present jobs quicker and in expanding the outlook of research work. Since activity in statistics necessitates knowledge of mathematics its use cannot be totally avoided, but it has been kept to a minimum.

Statistics is the study of measurements, involving collection, summarization and interpretation of data. And this is true in all instances, be it the businessman looking at his sales records, the dairy farmer contemplating the milk production of his cows or the state police recording causes of traffic accidents. It is also true of the scientist, of the social scientist studying child delinquency, of the biologist working with bacterial organisms or the agricultural scientist experimenting with animals and plants. All these people use the methods of statistics in one way or another, some requiring more advanced techniques than others. The diversity of their interests makes it impossible to demonstrate uses of a computer in all branches of statistics so the discussion is confined to some problems of the agricultural scientist whose work is familiar to the author.

Statistical analyses of experiments

One of the greatest uses of statistics in agriculture is in the interpretation of data obtained from experiments. These are usually designed to compare the effectiveness of different factors under the control of the experimenter, different plant varieties for example, different seeding prodecures, different fertilizer treatments, and so on. Let us consider a simple experiment that might be used for studying the value of four different fertilizers and three different varieties of a plant species,

* Biometrics Unit, Plant Breeding Department, Cornell University

corn perhaps. The data may be weights of corn taken from a field trial in which all 3 varieties had been planted in combination with each of the 4 fertilizers. Thus the actual layout of the experiment might be 12 plots in 3 rows of 4 plots each, one row to a variety and a column of 3 plots for each fertilizer. A diagram of the experimental plan would be as follows:

Variety of corn	Fertilizer			
	1	2	3	4
A				
B				
C				

In practice, an experiment of this nature would be planned a little differently, but the above suffices for our purposes. (See Federer, (5), for discussion of experimental designs).

After growing the plants and taking measurements the experimentalist asks the statistician "how do I tell if one variety is truly better than another? Variety A yielded 100 lb. more corn from its four plots than did variety B, but only 35 lb. more than variety C, does this mean A is greatly better than B and not much better than C?" The answers to questions such as these are based on comparing observed differences between varieties against measures of average variatility among all the plots. Generally speaking, calculating values of this average variability is one of the biggest computing tasks that statisticians have, not particularly so in this example but in data analysis problems generally, expecially those involving large volumes of data. It is therefore the point at which computers are of greatest assistance. The task is lengthy because it involves finding the square of each observation and of various sums of the observations, and adding these squares together in different ways. The result is what is called an Analysis of Variance table, which is a table of calculated values that attempts to summarize the variability between the observations according to its different sources. (Variance is the statistical term for variation,

or variability). The calculations for this experiment would be as follows:

Analysis of Variance

Source of variance	Calculation
Rows (varieties):	1/2 Sum over rows of (row mean - grand mean)^2
Columns (fertilizers):	1/3 Sum over columns of (column mean - grand mean)^2
Unexplained:	1/6 Sum over observations of (observation - row mean - column mean + grand mean)^2

In this case the analysis of variance table divides the variation among plots into variation due to variety differences, variation due to fertilizer differences - and residual variation which cannot be explained by either of the other causes. We need not be concerned here with interpreting this table but rather with factors involved in carrying out the required calculations. First of all, its constants 1/2, 1/3 and 1/6 are determined by the number of rows and columns in the experimental plan, being 1/(r-1), 1/(c-1) and 1/(r-1)(c-1) respectively, where r is the number of rows and c the number of columns. The computations that are lengthy are the obtaining of means, the taking of differences from them, squaring the differences and adding the squares. Although the form of the calculations shown in the above table is suitable for understanding their general meaning, it is not the most suitable for computing purposes. For example, the first line in the table indicates how the rows vary from their mean - but if calculated in this form rounding errors easily arise, for were a row total (of 4 observations) to be 23 should the mean be taken as 6, 5.7, or 5.8? Errors thus occurring can accumulate to serious proportions, but can be avoided by a fortunate algebraic simplification which is also computationally easier to use. Thus the calculation for the first line can be expressed as R - F where R = 1/8 Sum over rows of (row total)^2 and F = 1/24 (grand total)^2. These terms are easily computed and the problem of rounding errors arises only when making the divisions by 8 and 24. Even then one must ensure that sufficient significant digits are carried so that R - F is neither zero nor negative, for it is zero only when all row totals are the same and it is never negative. One might suggest that floating point arithmetic be used

to avoid rounding errors but it does not remove the problem entirely; for example, if R and F are both 12 - digit numbers with no decimals, in a 10 - digit floating point field requiring two digits for the exponent, then R - F in floating point arithmetic will be given as zero if its true value is less than 9999. Knowledge of one's own data and customary programming precaution will ensure that such errors do not occur. In practice the bulk of the calculating in this style of problem, the squaring and adding, is done in fixed point arithmetic; machine capacity seldom imposes a limit on the use of fixed point since the magnitude of individual observations is usually such that even several thousand can be processed satisfactorily, and the increased speed of fixed point arithmetic on most computers is then well worthwhile.

The calculations for the experiment described are simple and easily carried out; but not all such experiments are quite as straightforward. Had the experimentalist wanted analyses for several measurements on his corn instead of just weight, each analysis would involve calculations of the sort described. And if there had been 15 or 20 varieties of corn and 8 or 10 fertilizer treatments the computing for each analysis would be some fifteen times greater than in the experiment as outlined. Furthermore, if there had been not just one such experiment but several, at different times and places, the computing needs would be still greater, particularly if the number of varieties and fertilizers differed from one experiment to another. This sort of thing is not al all atypical of agricultural research. Sometimes too, research of a similar nature is conducted at the farmer level, where county extension agents help individual farmers with field trial plantings. In the author's home country of New Zealand, for example, approximately a thousand of these experiments are analysed annually in a central statistical office of the Department of Agriculture. Desk calculation and checking used to take half a day per experiment but now thirty minutes computer time sees 80 to 100 experiments processed.

Computers help the dairy farmer

Dairy farmers as a regular practice do not weigh the milk yield of each of their cows at the time they are milked. They therefore have no regular basis for comparing cows and deciding which are the better animals to keep for successive lactations and which to breed from. Those who join their

local Dairy Herd Improvement Association, however, have regular records made available to them on a monthly basis resulting from a technician's visit to the farm each month to take the necessary weighings, and to sample each cow's milk for determination of butterfat content. From each visit monthly production is estimated and over the year total lactation yield of each cow accumulated. These are the figures the farmer uses to make comparisons between cows, to compare a cow's record with her previous year's production and to compare cows with each other from month to month. These comparisons are of great value to the farmer in helping him improve the overall production of his herd, and having the necessary records is a service that has been available for many years based on desk calculations derived from the monthly weighings. Nowadays, however, computers are doing the work more and more and in many sections of the country giving the farmer not only production records but a great deal more besides. For example, for each cow the number of days since calving are accumulated and recorded, as is the expected date of the next calving; and also estimates of feed costs for each cow based on current feed prices, the cow's weight and her level of production; gross income from the milk produced and profitability over feed costs are also calculated. Each month these figures are sent to the farmer on specially-prepared forms within two or three days of the weighings being made on his farm, with figures not only for each cow but also totals and averages for the whole herd and comparisons with the herd summary of 12 months earlier. Each month this work is done for some 360,000 cows in New York State, Maryland, West Virginia and the New England states, using a computer in the Dairy Records Processing Laboratory of the Animal Husbandry Department of the New York State College of Agriculture at Cornell University, Ithaca, New York. Throughout the nation, in eleven other states, there is a similar use of computers handling in all, the records of over a million and a quarter cows each month. The popularity of these procedures is measured by their growth in recent years, see for example reference (4). Although the cost to the farmer in New York is approximately 8 cents per cow per month more through using a computer, the program centerred at Cornell has grown from 2,500 cows per month 5 years ago to its present capacity of 360,000 per month at the end of 1962. This is testimony to the value the farming community places on it as an aid to their improvement practices.

The records collected during the course of the activities just described are also valuable in another direction - as research data for those interested in statistical studies of the mode of inheritance of milk and butterfat production. In some species of both plants and animals the mode of inheritance of certain traits is well defined and the outcome of particular matings for producing a subsequent generation can be accurately predicted. But in other instances, such as production traits in farm animals, inheritance is undoubtedly a complex process and while the outcome of particular matings cannot be predicted with exactitude, certain characteristics of inheritance can be studied by appropriate statistical analyses of records of related animals. This is especially true nowadays in dairy production, as a result of the widespread use of artificial insemination, resulting in cows in many different herds being half-sisters through having the same sire. The production records of such animals serve two purposes. Firstly, by collecting together the records of groups of half-sisters, comparisons can be made among the sires used in an artificial breeding program. These comparisons enable an artificial breeding organization to maintain a team of bulls whose daughters are high yielding cows, by continually discarding those whose daughters productions are too low - and such comparisons, through being based on averages of many daughters in many herds, can be made very reliable. Here again computers aid the dairy farmer, for the task of searching through the records of thousands of cows in hundreds of herds to accumulate daughter totals (and hence averages) for many sires, is well beyond the practical scope of clerks and desk calculators. For example, in the Dairy Records Processing Laboratory at Cornell the records of approximately 7000 herds are used three times a year to assemble daughter averages of bulls used by the New York Artificial Breeders' Co-Operative, New York's largest artificial breeding organization. The records of more than 37,000 daughters of 156 sires were included in a recent summary, (1), an average of almost 240 per bull, one having as many as 3,852 daughters. That the average artificially-bred cow in New York outproduces her naturally-bred herd mates by some 300 pounds of milk per lactation, representing more than $8 million added farmer income per year, (9), is witness to the value of artificial breeding and to the usefulness of computers in their assisting with this work.

The second use of records of artificially-bred cows is that by their very relationship, being groups of half-sisters, studies can be made of variation in production among daughters of the same sire and of variation among sires as represented by their daughter groups. Studies of this kind are useful in learning about certain characteristics of the inheritance of milk production, see Searle, (11). They are undertaken by considering the data (records of daughters of different sires, in different herds) in the same way as discussed earlier for the experiment with corn varieties and fertilizer treatments. The rows now represent herds and the columns represent sires and the statistical analysis involves apportioning the variance (variation) among the records according to various sources, namely that arising from herd differences, sire differences and other causes. Since not every sire used in artificial breeding has a daughter in every herd that uses the service, not all cells of the grid have a record in them - and likewise where a sire has more than one daughter in a particular herd there will be more than one record in that herd-sire cell. The mathematics of this situation need not be considered here (see Searle, 10) suffice to say that it results in large amounts of data being needed for the calculated variances to be reliable estimates of the underlying variation among all such records. Since the arithmetic involved is tedious, being sums of squares more complex than those considered earlier, the calculations are ideally suited to computers. Studies of perhaps five or six hundred cows from ten or a dozen sires in some 60 herds (i.e. 60 rows and 10 columns in the table of data) were once thought respectable, in terms of size, but nowadays ten thousand records and more, in a table of possibly 800 rows (herds) and 150 columns (sires) are considered acceptable, the general criterion being the more data the better.

Simulation technics

Statistical procedures make great use of sampling. For example, instead of calculating the true mean of a population, such as the mean height of all people in New York State, we can use just a sample of the population and from the mean of the sample make inferences about the mean of the population. The value of the sample mean as an indicator of the population mean depends on the size of the sample as well as the manner in which it is drawn. Different samples give different sample means, a whole range of which gives more information about the population mean than does a single one. The

same is true of the genetics problem just discussed, of estimating the variance among production records of dairy cows born from artificial breeding. one set of data gives one set of estimates and further data give other estimates. In the case of estimating means we can make quite good inferences about the population mean from a single sample mean because the statistical properties of the situation are usually relatively clear. But the situation is not at all well known when estimating variances from data such as just discussed. One way to learn something about the properties of sample variances in these situations is to calculate them for large numbers of samples of data. Even then we do not know the true population variances and cannot make comparisons easily between different methods of calculating the sample variances, (see Henderson, (8), for example) This is where computers help - they can be used to generate samples from hypothetical populations having known variances; from each sample estimated variances are calculated, and over a large number of samples the range and distribution of the estimates is compared with population variances put into the hypothetical population generated by the computer from which the samples were derived. In this way several valuable comparisons can be made. First, from a population with known variances, random samples of the same size and pattern can be drawn and their estimated variances plotted against the population variances. This should tell us something about the statistical procedures involved when the type of sample and population is held constant. Secondly, the process can be carried out for population variances of different magnitudes to see how this affects the estimates; it can also be carried out for samples of different size from the same population, to study the effects of sample size on the distribution of estimated variances. And the procedures can also be undertaken for different methods of calculating the estimated variances to investigate possible differences between them. All this requires a great deal of computing, for samples large in size and many of them in each case would be envisaged. For example, one might take 500 samples of 10,000 from each population specified to the computer - and for each sample the computer has first of all to construct the sample from the population and then calculate the estimated variances. In this way we are simulating populations, drawing samples from them and conducting statistical procedures thereon in order to learn something about these procedures, a process similar to that of the traffic engineer who simulates

traffic flows for a large city to learn something about road construction plans. Some results of using these simulation techniques in variance estimation have recently been reported by Bush and Anderson, (3), and in estimating genetic correlations by Van Vleck and Henderson, (12).

Simulation procedures have also been used to study variables involved in the inheritance mechanism itself. In particular, studies have been made of the process known as selection, where improvement of a species for a particular trait is sought by selecting as parents for the next generation those individuals which display superiority in that trait in their own generation. For example, a dairy farmer is continually discarding his lower-producing cows not only because he wants only high-producing animals in his herd but also he does not want to breed succeeding generations from low-yielding cows. In this way he improves his herd. Improvement in this context generally relates to improvement in genetic merit. Studies of achieving this through selection can be made on a computer by simulating a group of individuals in the computer, giving values to the hypothetical genetic elements involved, simulating matings and the inheritance mechanism to evolve a new generation, adding hypothetical environmental effects to obtain expressions of the trait, and simulating selection on the basis of these to choose individuals to act as parents for the next matings. The point of interest is to plot the extent to which this selection process improves the genetic merit of the population for the trait concerned. Although simulation cannot include all the factors that affect this selection process in real life it is useful in giving guidance as to how some of them operate in hypothetical situations and therefore how they may operate in actual practice. One great advantage is the computer's ability to "breed" at high speed and create successive generations. Let us consider a very simple example of how the technique operates.

Heredity is basically governed by units of inheritance called genes, which occur in pairs, one or more pairs conditioning the expression of a trait. When two individuals are mated each gives one gene of each of his own gene pairs to the newly-formed individual - exactly which one being a random event. Thus if a trait is governed by a single pair of genes which we will denote by A or a, an individual will be AA, aa, or Aa. At mating,

no matter which one of these pairs an individual has, one randomly selected gene of the pair will be transmitted to the embryonic individual; and when a trait is governed by more than one gene pair the process applies independently for each pair.

To demonstrate the simulation technique we will suppose for simplicity's sake that we are dealing with a trait controlled by two pairs of genes, A, a and B, b. Assume we start with an initial population of two males m_1 and m_2 having genes AAbb and aaBB respectively, and two females f_1 and f_2 with genes AaBb and aabb. Let us give hypothetical values to the genes of 1 for A and B and 0 for a and b and we will suppose that the gene value of an individual is obtained by simple addition of the values for each gene separately. Then the genetic values of the initial population are as shown in Table 1. We will consider a breeding situation of six matings,

(Show Table 1)

randomly selecting one male and one female to give one offspring from each mating. To the genetic value of each offspring is added a random term representing environmental effects to give a value for the expression of the trait by the offspring; and on the basis of these values (genetic plus environmental) four offspring, the best two of each sex, are selected to be parents for the next generation. And so the simulation proceeds for successive generations, progress in the mean genetic value being plotted for each generation. Table 2 shows the various steps in the procedure.

(Show Table 2)

The first column sets out six matings supposedly made at random among the initial population of Table 1; columns (2) and (3) show random genes, one gene per gene pair, from each parent. Sex is given arbitrarily to the offspring in column (4), retaining a 1:1 ratio, and the genetic composition of the offspring is shown in columns (5) and (6), derived from columns (2) and (3) using the genetic values already decided on. Column (7) represents the random environmental terms, and column (8) is the sum of these and the genetic values, this sum representing the actual expression of the trait. On these values the two highest males and females are selected as parents of the next generation; those selected are indicated by an asterisk in columns (9) and (10) and their genetic composition as derived in column (5)

is shown again in column (11). These four individuals are now used as parents of the next generation repeating the whole process from column (1), and so on for as many generations as we wish, computer time permitting.

This is a small and simplified example of how genetic processes can be simulated on a computer. Many other factors can be included and of course, the magnitude of the whole situation greatly expanded, to 20 gene pairs say, instead of 2, to various initial gene frequencies (in this case 3A to 5a and 3B to 5b, in Table 1), to types of gene values different from those used here, including interaction effects, and also for different numbers of matings and selections. Genetically the interest lies in seeing how variations in these factors affect the mean genetic values obtained at each generation. Results of such work have recently been reported by Fraser, (6) and (7), and Barber (2). They are of great interest in indicating some of the effects which these factors have on selection, a procedure that is widely used for species improvement.

Conclusion

Additional illustrations of computer uses in statistics and genetics could well be found, but those already discussed present many of the salient features of both practical applications and research activities in these fields. Both areas of work contain instances of data processing ideally suited to using a computer and of situations where simulation techniques are yielding information that was practically unobtainable before the advent of high speed computers.

References

(1). A. I. Daughter Level Report, Animal Husbandry Dept., Cornell University, Ithaca, N. Y., September, 1962.

(2). Barber, J. S. F. Simulation of genetic systems by automatic digital computers. III and IV. Aust. J. Biol. Sciences 11, 603-625, 1958

(3). Bush, Norman and R. L. Anderson Estimating variance components in a multi-way classification. Mimeo Series No. 324, Institute of Statistics, Consolidated University of North Carolina, 1962.

(4). Dairy Herd Improvement Letter, Agricultural Research Service, 44-114, of the National Co-operative Dairy Herd Improvement Program, U. S. Dept. of Agriculture, Washington, D.C., 1962.

(5). Federer, W. T. "Experimental Design" The MacMillan Company, N. Y., 1955.

(6). Fraser, A. S. Simulation of genetic systems by automatic digital computers. I and II. Aust. J. Biol. Sciences 10, 484-499, 1957.

(7). Fraser, A. S. Simulation of genetic systems by automatic digit computers V. "Biometrical Genetics" ed. Oscar Kempthorne, Pergamon Press, N. Y., 70-83, 1960.

(8). Henderson, C. R. Estimation of variance and covariance components. Biometrics 9, 226-252, 1953.

(9). Rockefeller, Nelson A. Dedicatory Address, Dedication Ceremony of Frank B. Morrison Hall, N. Y. State College of Agriculture, Ithaca, N. Y., 1962.

(10). Searle, S. R. Sampling variances of estimates of components of variance. Ann. Math. Stat. 29, 167-178, 1958.

(11). Searle, S. R. Estimating the heritability of butterfat production. J. Agr. Sci. 57, 289-294, 1961.

(12). Van Vleck, L. D. and C. R. Henderson Empirical sampling estimates of genetic correlations. Biometrics 17, 359-371, 1961.

TABLE 1

Initial Population in a Simulation Process

Sex	Identity	Genetic Composition Symbol	Value
Males	n_1	AAbb	2
	m_2	aaBB	2
Females	f_1	AaBb	2
	f_2	aabb	0
Mean genetic value			1.5

TABLE 2

Simulation of Random Matings and Selection

Random matings	Random genes from each parent		Offspring			Random environmental term	Expression of trait	Selection of parents		
				Genetic composition				Top 2		
	Male	Female	Sex	Symbol	Value			Males	Females	Genetic composition (from col.5)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
m_1f_1	Ab	AB	m	AABb	3	3	6	*		AABb
m_1f_2	Ab	ab	f	Aabb	1	4	5		*	Aabb
m_2f_1	aB	aB	m	aaBB	2	1	3			
m_2f_2	aB	ab	f	aaBb	1	3	4		*	aaBb
m_1f_1	Ab	Ab	m	AAbb	2	4	6	*		AAbb
m_2f_1	aB	aB	f	aaBB	2	1	3			
			Mean genetic value		1.83					